THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 266, No. 2, Issue of January 15, pp. 1229-1236, 1991
Printed in U.S.A.

Cloning of *Amb a I* (Antigen E), the Major Allergen Family of Short Ragweed Pollen*

(Received for publication, June 22, 1990)

Thorunn Rafnar‡, Irwin J. Griffith§, Mei-chang Kuo§, Julian F. Bond§, Bruce L. Rogers§, and David G. Klapper‡¶

From the ‡Department of Microbiology and Immunology, University of North Carolina School of Medicine, Chapel Hill, North Carolina 27599 and the §ImmuLogic Pharmaceutical Corporation, Cambridge, Massachusetts 02139

To determine the structure of *Amb a I* (previously called antigen E), the major allergen from short ragweed, cDNA from pollen was cloned into λgt11 and λgt10. One of the three distinct clones isolated from the λgt11 library by screening with anti-denatured *Amb a I* antibodies was used to screen both libraries for other *Amb a I* sequences. Multiple clones were isolated and sequenced and proved to be highly homologous but nonidentical. The clones could be divided into three groups based on sequence similarity, and in accordance with the International Union of Immunological Societies-approved nomenclature (Marsh, D. G., Goodfriend, L., King, T. P., Lowenstein, H., and Platts-Mills, T. A. E. (1986) *Bull. WHO* 64, 767–770) they have been designated *Amb a I.1*, *Amb a I.2*, and *Amb a I.3*. Clones within a group have greater than 99% identity, and similarity among groups is 85–90% at the nucleotide level. The amino acid sequence of four peptides (isolated from antigen E obtained from the Research Resources Branch of the National Institutes of Health) containing 132 amino acids was identical to one of the clones (*Amb a I.1*). The presence of multiple naturally occurring isoelectric forms of *Amb a I* was demonstrated by two-dimensional gel electrophoresis and Western blotting. Southern blot analysis demonstrates the presence of multiple *Amb a I*-related sequences in the ragweed genome. *Amb a I* is therefore not a single molecule but rather a family of closely related proteins.

Of all the seasonal aeroallergens, pollen from short ragweed (*Ambrosia artemisiifolia*) is perhaps the most clinically important. The extremely small particle size allows this pollen to be carried hundreds of miles, and it is the major cause of late summer hay fever in the eastern United States and Canada (2). Of the 52 antigens present in an aqueous extract of pollen, at least 22 are allergens defined by their reactivity with human IgE (3). At least five significant human allergens from short ragweed pollen have been purified to homogeneity and studied with respect to their biochemical and immunological characteristics (2).

Antigen E or *Amb a I* (according to nomenclature in Ref. 1) is considered the most important allergen since 95% of ragweed-sensitive individuals react to it in skin tests and show high IgE antibody titers to it (4, 5). *Amb a I* is highly abundant, comprising about 6% of the total protein in a neutral aqueous extract of pollen (4). It is an acidic, amino-terminal "blocked," reportedly nonglycosylated single-chain protein of 38 kDa which undergoes proteolysis during chromatographic purification and is cleaved into two chains, α and β, of 26 and 12 kDa, respectively (6). The two-chain form is reported to be allergenically and antigenically indistinguishable from the intact molecule, but modification of the protein, including reduction and alkylation of disulfide bonds, urea denaturation and renaturation, or succinylation of lysine residues, reduces the IgE immunoreactivity of the molecule (2). Recent data with Western blotting demonstrate that *Amb a I* retains the ability to bind antibody from allergic humans and hyperimmunized animals (documented in Fig. 3 and Footnote 1).

It has been reported that immunotherapy utilizing purified *Amb a I* is as effective in alleviating clinical symptoms in allergic patients as is immunotherapy using whole pollen extract (7). Immunizations with modified forms of *Amb a I* have also been tested and shown to be as clinically effective as native *Amb a I* and to cause fewer systemic reactions (8–11). *Amb a I* has been shown to have three nonoverlapping, nonrepeated antigenic sites, as defined by murine monoclonal antibodies, of which at least two represent major human allergenic epitopes (12). In addition, preliminary studies directed at examining T cell epitopes of *Amb a I* suggest that they are linear rather than conformational (13–15).

At present, desensitization immunotherapy for ragweed-allergic individuals relies upon multiple injections of small doses of aqueous pollen extracts. These protocols are not ideal since individuals present with varying sensitivities to each of the multiple components in an extract, and various batches of extract used for diagnostic and therapeutic purposes vary a great deal in their specific allergen content. Furthermore, although immunotherapy offers some improvement to many patients, almost no patients become completely asymptomatic, and a number of patients show no symptomatic improvement at all (16). Identification and characterization of specific epitopes of *Amb a I* might be particularly useful in improving an immunotherapeutic approach to desensitization.

As a first step toward this end, cDNA libraries have been constructed from short ragweed pollen and whole flowers. This report describes the cloning of *Amb a I* and presents the complete nucleotide and deduced amino acid sequences of three clones coding for this 398-amino acid allergen. Comparison of the three cloned sequences shows that they have

* This work was supported by National Institutes of Health Grant AI 14908 (to D. G. K.) and by a grant from the North Carolina Biotechnology Center (to D. G. K.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

¶ To whom correspondence should be addressed.

[1] T. Rafnar, I. J. Griffith, M-C. Kuo, J. F. Bond, B. L. Rogers, and D. G. Klapper, unpublished results.

several differences, suggesting that *Amb a I* is actually a family of proteins with at least three members. These proteins represent an as yet undescribed family since a search of available data bases shows no significant homology to other proteins from either procaryotes or eucaryotes. Finally, in contrast to what has been reported in the literature (6), the carboxyl terminus of *Amb a I* appears to be cysteine. During this investigation, a peptide corresponding to the predicted carboxyl terminus of *Amb a I* was isolated from chromatographically purified *Amb a I*, confirming the assignment of cysteine as the carboxyl-terminal amino acid.

MATERIALS AND METHODS

Antibodies—Mouse monoclonal anti-denatured *Amb a I* antibodies JB4F3-5, JB3C9-3, JB1E3-4, 2D8/E6, JB4E3-3, JB4E6-4, and IB1E2-2 have been described previously (17).

Plant Tissue—Flowers and leaves of short ragweed were picked, frozen immediately in liquid nitrogen, and stored at −70 °C until processing. Defatted short ragweed pollen and meadow fescue pollen were obtained from Greer Laboratories (Lenoir, NC).

Genomic DNA Isolation—Genomic DNA was isolated by a method published previously (18) with the following changes. 50–100 g of frozen ragweed flowers were ground in liquid nitrogen using a mortar and pestle. The dry powder was suspended in homogenization buffer (0.2 M Tris-HCl, pH 8.5, 0.2 M sucrose, 6 mM KCl, 50 mM $MgCl_2$, 5.8 mM 2-mercaptoethanol, 2% (w/v) polyvinylpyrrolidone 40 and homogenized at high speed with a Polytron homogenizer (Brinkmann Instruments) for 3 min. The DNA was banded two times on CsCl gradients to obtain DNA that could be digested by restriction enzymes (21).

RNA Isolation—RNA was isolated from 10 g of defatted ragweed pollen or whole flowers by a slightly modified method published by Lagrimini *et al.* (19). After the original phase separation, the nucleic acids were precipitated in the presence of 0.3 M sodium acetate and 75% ethanol at −20 °C for at least 4 h. The precipitate was pelleted by centrifugation at 13,000 × *g* for 20 min and the pellet dissolved in 2 ml of sterile water. The RNA was then precipitated selectively in the presence of 3 M LiCl at −20 °C overnight. The precipitate was pelleted as before.

The RNA pellet was dissolved in 1 ml of oligo(dT) binding buffer (10 mM Tris-HCl, pH 7.5, 0.5 M NaCl, 0.5% SDS,[2] 1 mM EDTA) and poly(A^+) RNA selected on oligo(dT)-cellulose (Collaborative Research). 60 μg of hydrated oligo(dT)-cellulose was added to the sample, heated to 55 °C for 1 min, and mixed gently at room temperature for 10 min. The resin was centrifuged for 5 s in a microcentrifuge, the supernatant poured off, and the sample washed three times with fresh binding buffer before being eluted with 2 × 300 μl of elution buffer (10 mM Tris, pH 7.5, 0.5% SDS, 1 mM EDTA). The poly(A^+) RNA was precipitated with 0.01 volume of 3 M sodium acetate and 2.5 volumes of ethanol.

Isolation and Characterization of cDNA Clones—A λgt11 library was generated from cDNA made from pollen mRNA using avian myeloblastosis virus reverse transcriptase (Life Sciences) and RNase H (Boehringer Mannheim) as described (20). After methylation with *Eco*RI methylase (Bethesda Research Laboratories), cDNA was ligated with phosphorylated *Eco*RI linkers (Bethesda Research Laboratories), digested with *Eco*RI, and excess linker removed from the cDNA on a Sephadex G-100 column (Pharmacia LKB Biotechnology Inc.). The cDNA was ligated into dephosphorylated λgt11 arms (Promega Biotech, Madison, WI), packaged with the Gigapack packaging system (Stratagene Cloning Systems), and plated using standard procedures (21). The library was plated on 150-mm Petri dishes at a density of 20,000–30,000 plaques/plate and screened with a pool of monoclonal mouse anti-*Amb a I* antibodies using the Protoblot kit (Promega Biotech). Antibodies from ascites fluids were prepared by preabsorption with filters of nonrecombinant λ plaque lifts and used at a dilution of 1:5000. Three antibody-binding clones were purified by repeated plating (two to four times), subcloned into M13mp19, and sequenced using the dideoxy sequencing method (22). One of these clones, *Amb a I.2*, was radiolabeled and used to rescreen the library for related nucleotide sequences. This screening yielded six independent additional clones.

A λgt10 library containing cDNA inserts from short ragweed flower head mRNA was prepared according to published procedures (23, 24) using an Amersham Corp. cDNA cloning kit. Approximately 8 × 10^4 independent clones were screened by standard procedures (25) using nick-translated (26) ^{32}P-labeled *Amb a I.2* cDNA as a probe. Four independent hybridizing cDNA clones were isolated and subcloned into M13 for dideoxynucleotide sequencing (22).

Northern Blotting and Hybridizations—20 μg of total RNA was electrophoresed through a formaldehyde-containing gel, blotted onto nitrocellulose paper, and hybridized to a radiolabeled cDNA clone according to standard procedures (21). Posthybridization washes were done in 2 × SSC, 0.1% SDS, first at room temperature for 15 min, and then for 1 h at 50 °C. The filter was air dried and exposed to Kodak XAR film at −70 °C with an intensifying screen.

Southern Blot Analysis—Digestion of genomic DNA, agarose gel electrophoresis, and blotting onto nitrocellulose were done by conventional methods (21). The filters were hybridized to nick-translated cDNA clones. Hybridization and washing conditions were the same as those used for RNA blots.

Protein Purification and Sequencing—*Amb a I* protein sequence data were obtained from one chromatographic form of *Amb a I* (AgE-B) generously provided by Dr. T. P. King (Rockefeller University, New York) and from the *Amb a I* protein, which was purified according to the method of King *et al.* (6) with certain modifications. In brief, the defatted short ragweed pollen (Greer Laboratories) was extracted in 50 mM Tris, pH 8.0, containing 0.1 mM phenylmethylsulfonyl fluoride, 1 μg/ml leupeptin, 10 μg/ml pepstatin, and 10 μg/ml soybean trypsin inhibitor. After decolorization with DE52 cellulose (Whatman) the extract was fractionated by ammonium sulfate precipitation. The 45–59% saturation fraction was purified further by a Sephacryl S-200 (Pharmacia) column and Mono Q (Pharmacia) anion-exchange column. The *Amb a I* protein purified by this modified method appeared as a single band of 38 kDa on SDS-polyacrylamide gel electrophoresis (SDS-PAGE). The majority of amino acid sequence data was derived from mixtures of peptides produced either by cyanogen bromide (CNBr) or formic acid (29) cleavage of the AgE-B or *Amb a I*. The peptide mixtures were treated in the sequenator with *o*-phthalaldehyde to block all primary amines when a proline was known to be present at the amino terminus (30), which was identified in an exploratory analysis of the peptide mixtures. One of the peptide mixtures derived from a CNBr digest and an aliquot of AgE-B was subjected to SDS-PAGE. The SDS-PAGE was run on a precast 10–20% Tricine minigel (Novex) run at 100 volts for 1.5 h. Proteins were transferred to polyvinylidene difluoride membrane (Millipore) in 10 mM CAPS buffer, pH 11, at 150 mA for 2 h (27). The polyvinylidene difluoride was stained with Coomassie Blue. The 12-kDa β chain band from AgE-B (6) and a 10-kDa band from the CNBr digest were cut out for protein sequencing. Protein sequencing was performed on an Applied Biosystems model 477A gas-phase sequenator with on-line phenylthiohydantoin derivative analysis (model 120). The protein was alkylated *in situ* in the sequenator by using the nonnucleophilic reductant, tributylphosphine, with concomitant alkylation by 4-vinylpyridine in ethylmorpholine buffer (28).

Two-dimensional Western Blot—Isoelectric focusing was performed on a Hoefer gel apparatus with 15 μg of crude soluble pollen protein. The gel consisted of 7.5% acrylamide with 3.5% Pharmalytes, pH 4.5–5.3 (Pharmacia) and 3.5% Ampholines, pH 3.5–10 (LKB), run at 13 watts for 3.5 h until a constant voltage was reached (31). The gel section was placed on a slab of 10% acrylamide SDS-PAGE and electrophoresed for 3.5 h at 40 mA according to the protocol cited (32). The proteins were transferred overnight in phosphate buffer (20 mM $NaPO_4/NaH_2PO_4$, pH 6.8) to 0.1-μm nitrocellulose (Schleicher & Schuell) at 0.2 A (33). The blot was rinsed in blot solution (25 mM Tris-HCl, pH 7.5, 0.171 M NaCl, 0.05% Tween 20; Sigma). The first antibody incubation was overnight at room temperature with a 1:1000 dilution of goat anti-*Amb a I* IgG in blot solution. The excess first antibody was removed with three 15-min rinses with blot solution. The second antibody was a 1:2500 dilution of biotinylated swine anti-goat IgG (Boehringer Mannheim) in blot solution for 2 h. The blot was then rinsed with blot solution three times (15 min each) and incubated for 1 h in blot solution with ^{125}I-streptavidin (2 μCi, Amersham Corp.). The blot was rinsed with blot solution until the waste wash returned to background. The blot was then exposed to film at −80 °C overnight.

Sequence-specific Amplification of Amb a I cDNA with PCR—4 g of defatted pollen (Greer Laboratories) was ground with a mortar and

[2] The abbreviations used are: SDS, sodium dodecyl sulfate; PAGE, polyacrylamide gel electrophoresis; Tricine, *N*-[2-hydroxy-1,1-bis(hydroxymethyl)ethyl]glycine; CAPS, 3-[cyclohexylamino]-1-propanesulfonic acid; PCR, polymerase chain reaction.

pestle in 4 M guanidinium isothiocyanate buffer and total RNA isolated by standard procedures (34).

Priming oligonucleotides (Table I) for DNA amplification (35) were RW38, RW32, RW45, and anchor primer. RW38 corresponded to the amino-terminal coding strand sequence encoding amino acids Leu-Tyr-Phe-Thr-Leu (amino acids 10–14). The RW38 sequence was conserved between *Amb a I.2* and *Amb a I.3*. RW32, which corresponded to the noncoding strand sequence between 12 and 39 nucleotides 3′ of the TAA stop codon, was specific to *Amb a I.1*. RW45 corresponded to the noncoding strand sequence complementary to amino acids Ile-Lys-Ser-Asn-Asp-Gly (amino acids 181–186) of *Amb a I*. Two additional primers, anchor template and anchor linker, were used for linking to cDNA. The anchor primer oligomer sequence was contained within the anchor template sequence. The anchor linker oligomer was phosphorylated (see Table I for specific sequences and restriction sites). These oligonucleotides were purchased from Research Genetics (Huntsville, AL).

First-strand cDNA was synthesized from 1 μg of total RNA with the cDNA Synthesis System Plus kit (Amersham Corp.) using poly(dT) as a primer. The single-stranded DNA (20 μl) was mixed with 100 pmol of each priming oligo, RW38 and RW32, 10 μl of 10 × reaction buffer (GeneAmp kit, U. S. Biochemicals), and 0.5 μl of *Thermus aquaticus* polymerase (U. S. Biochemicals). The mixture was brought to 100 μl with distilled water and overlayed with mineral oil (Sigma). The sample was amplified with a programmable thermal controller from MJ Research, Inc. (Cambridge, MA). The first five rounds of amplification consisted of denaturation at 94 °C for 1 min, annealing of primers to the template at 45 °C for 1.5 min, and chain elongation at 70 °C for 4 min. The final 20 rounds of amplification consisted of denaturation as above, annealing at 55 °C for 1.5 min, and elongation as above.

Amplified DNA was recovered by sequential chloroform, phenol and chloroform extractions followed by overnight precipitation at 4 °C with 0.5 volume of 7.5 M ammonium acetate and 1.5 volumes of isopropyl alcohol. DNA was digested simultaneously with *Eco*RI and *Pst*I and electrophoresed on a 1% GTG agarose (FMC, Rockland, ME) preparative gel. The predicted 1.2-kilobase band was isolated and recovered by glass bead adherence (Geneclean kit, BI0101, La Jolla, CA). The digested DNA was ligated into *Eco*RI/*Pst*I-digested M13 for dideoxy sequencing (22).

The very 5′ end sequence of *Amb a I* was determined using a modification of the anchored PCR (36). Double-stranded cDNA was synthesized from 1 μg of RNA with the cDNA Synthesis System Plus kit using poly(dT) as a primer, blunt ended with T4 polymerase, and blunt end ligated to self-annealed anchor template and anchor linker primers. Linked cDNA (3 μl) was mixed with 100 pmol of the anchor primer and RW32 primers, 10 μl of 10 × reaction buffer and 0.5 μl of *T. aquaticus* polymerase. The mixture was brought to 100 μl and amplified as described above. 1% of the volume of the primary PCR was reamplified with anchor primer and RW45 oligomers. RW45 is nested (internal) relative to oligomer RW32 used in the primary PCR. Amplified DNA from the secondary PCR corresponding to the 5′ end RW45 sequence was recovered as above, digested with *Kpn*I, and ligated into *Kpn*I/*Hinc*II-digested M13 for dideoxy sequencing (22).

RESULTS

Three clones were isolated from a λgt11 ragweed pollen cDNA library by screening with a pool of seven mouse monoclonal antibodies raised to denatured *Amb a I*. The sequences of all three of these clones showed extensive homology to peptides isolated from highly purified *Amb a I*. One of these clones (*Amb a I.2*) was radiolabeled and used to screen this library and a λgt10 library constructed from cDNA made from whole ragweed flowers, resulting in the isolation of 10 additional clones.

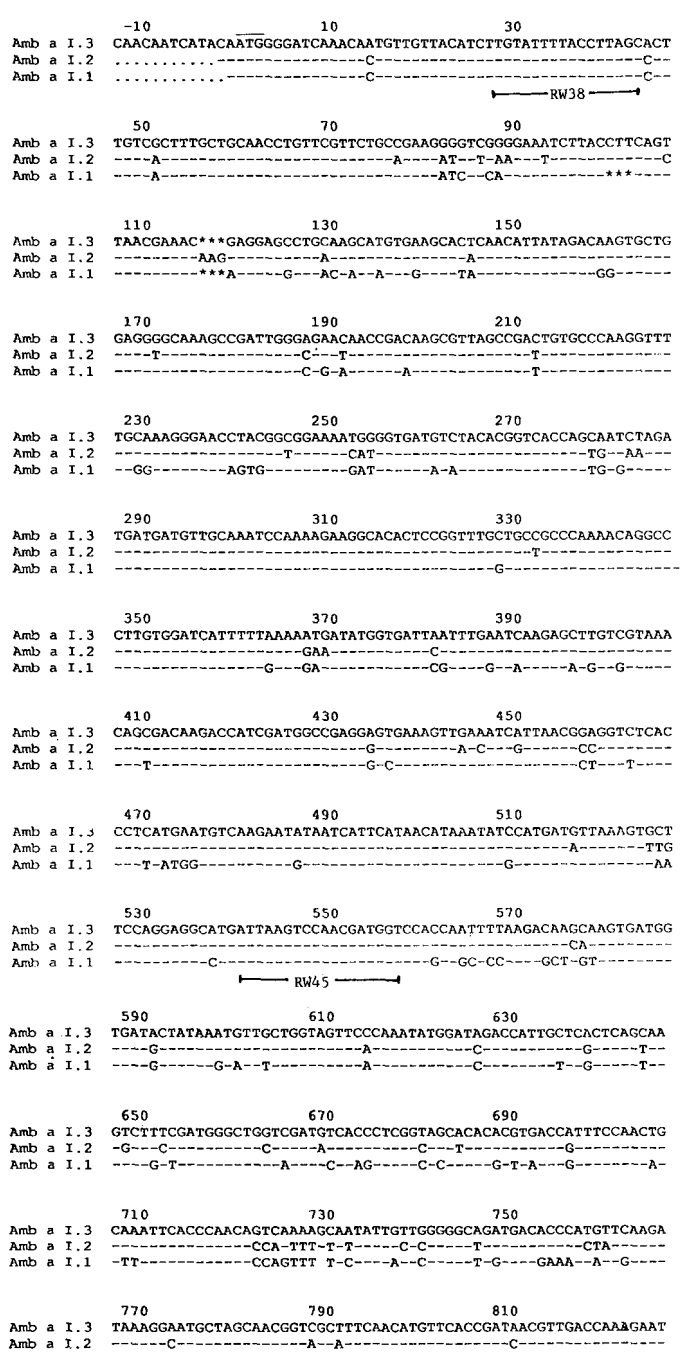
All the clones sequenced could be divided into three groups, *Amb a I.1, Amb a I.2,* and *Amb a I.3,* in which clones within a group share greater than 99% identity, and identity among groups ranges from 85 to 90% at the nucleotide level. The DNA sequences of the three largest prototypic clones in each group are shown in Fig. 1. The longest clone, a member of the *Amb a I.3* group, is 1331 bases long, has 13 nucleotides preceding the putative ATG start codon at position 1, an unbroken reading frame of 1190 nucleotides ending with a *TAA* stop codon, and a 125-nucleotide untranslated AT-rich region before the poly(A) addition site. The other clones start at positions −2 and 44 relative to the *Amb a I.3* prototype start site and have unbroken reading frames extending to a stop site identical to that found in the *Amb a I.3* group. The traditional mammalian consensus sequence for polyadenylation, AATAAA, does not occur in the untranslated regions. The *Amb a I.2* group has a nucleotide triplet at position 118 which is not present in the *Amb a I.3* and *Amb a I.1* groups, and the *Amb a I.1* group lacks an additional triplet at position 101 relative to the other two groups of clones. Since only a single non-full-length clone representing the *Amb a I.1* group was originally isolated from the libraries, the full-length clone was obtained by sequence-specific amplifications of pollen RNA using Taq polymerase and synthetic DNA primers. The amino-terminal primer corresponded to coding sequence nucleotides 25–44 in clone *Amb a I.3* and was conserved between the *Amb a I.2* and *Amb a I.3* groups. The carboxyl-terminal primer was specific to *Amb a I.1* and corresponded to the 3′-noncoding sequence between 12 and 29 nucleotides 3′ of the TAA stop codon in that particular clone. The very 5′ end sequence of *Amb a I.1* was determined using anchored PCR methodology. In this case, the amino-terminal primer corresponded to a synthetic linker sequence. The carboxyl-terminal primer described above was used in a primary amplification whereas a carboxyl-terminal primer corresponding to the noncoding strand sequence for amino acids 181–186 was used in a secondary amplification. A secondary amplification using a nested primer was necessary since the primary amplification

TABLE I

Nucleotide sequence of oligomers used in the PCR

RW32, RW38, and RW45 correspond to the coding or noncoding strand sequence of *Amb a I*. RW32 is *Amb a I.1* specific. Anchor primer (AP), anchor linker (AL), and anchor template (AT) do not correspond to *Amb a I* sequences. Nucleotide numbers correspond to those in Fig. 1. Restriction sites are underlined.

Oligomer	Sequence	Strand	Nucleotides
RW32	5′ GGGCTGCAGTCATTATAAGTGCTTAGT *Pst*I	Noncoding	1211–1228
RW38	5′ GGGAATTCTTGTATTTTACCTTAGC *Eco*RI	Coding	27–43
RW45	5′ ACCATCGTTGGACTTAAT	Noncoding	540–557
AL	5′ p-AATGATCGATGCT *Cla*I		
AP	5′ GGGTCTAGAGGTACCGTCCG *Xba*I *Kpn*I		
AT	5′ GGGTCTAGAGGTACCGTCCGATCGATCATT *Xba*I *Kpn*I *Cla*I		



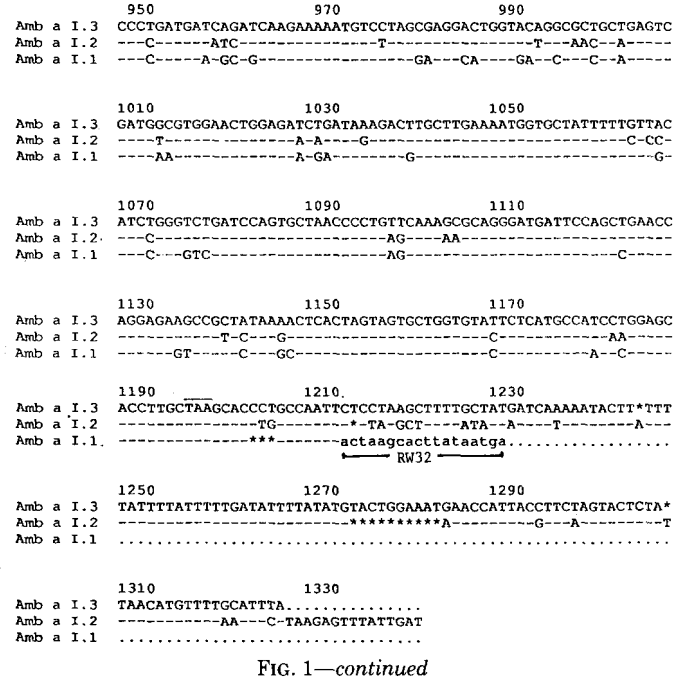


FIG. 1—*continued*

FIG. 1. **Nucleotide sequences of short ragweed clones *Amb a I.3, Amb a I.2,* and *Amb a I.1*.** *Numbering* begins with the adenosine of the ATG initiation codon of clone *Amb a I.3*. The start codon and TAA stop codon are *overscored. Horizontal bars* (- - -) represent identity to clone *Amb a I.1*; a *dotted line* (.....) indicates no sequence information; *asterisks* (*) mark positions at which nucleotides are absent when compared with the other sequences. *Lower case letters* 3′ in the clone *Amb a I.1* indicate the sequence of the 3′ oligonucleotide used for its amplification. The poly(A) tails of the clones are not shown. *Solid lines* (——) indicate the positions of the oligomers used in the PCR.

did not yield a discrete product band. Sequence identities at the nucleotide level in the coding regions are 91% between clones *Amb a I.3* and *Amb a I.2* and 86% between *Amb a I.2* and *Amb a I.1* and between *Amb a I.3* and *Amb a I.1*.

Amino acid sequence data are presented in Fig. 2. Sequences derived from chromatographically purified *Amb a I* accompany translated amino acid sequences of the three prototype clones. Approximately 130 amino acids were identified from four peptides. Two of the peptides were sequenced after a cyanogen bromide cleavage of purified *Amb a I*, and it was assumed that each follows a methionine. A third peptide was sequenced after acid cleavage (70% formic acid) of the aspartic acid-proline bond at position 361. For peptides 277–307 and 361–398, *o*-phthalaldehyde was used to block all the potential amino-terminal amino acids other than proline before the first sequencing cycle. For peptides 171–186, the *o*-phthalaldehyde blocking was applied before the seventh sequencing cycle to obtain the unique sequence between 177 and 186. The NH_2-terminal end of this peptide was obtained from the 10-kDa band that was isolated after SDS-PAGE. The fourth sequence was derived from the 12-kDa β chain band of AgE-B. Judging from the yield of protein sequencing analysis, the peptide sequence 45–91 was derived from a minor component of the β chain band. Thus, 129 amino acids or approximately one-third of the coding sequence for *Amb a I* has been obtained directly from purified protein. Also in Fig. 2 are the translated sequences for each of the three prototype clones coding for *Amb a I*. The proteins coded for by these sequences are 396–398 amino acids long and each has a clear hydrophobic segment following the initiating methionine. Each group has a single potential glycosylation site at position 37 relative to the initiating methionine. The clone representing group *Amb a I.1* corresponds most precisely to known amino acid sequences of peptides derived from enzymatic and chemical cleavages of chromatographically isolated *Amb a I*. The *Amb a I.3* and *Amb a I.2* clones differ by as much as 20% from the known amino acid sequence. The deduced amino acid sequences of clones *Amb a I.3* and *Amb a I.2* are 86% identical to each other (92% similarity), and each has approximately 76% identity (85% similarity) to clone *Amb a I.1*. Nucleotide and amino acid sequence comparisons and calculations were

```
                10                  30                  50
Amb a I.3  MGIKQCCYILYFTLALVALLQPVRSAEGVGEILPSVNET*RSLQACEALNIIDKCWRGKA
Amb a I.2  ----H----------T---------D-E-F---A---R---K----H--------C--
Amb a I.1  ----H----------T---------DLQ----*----*-R-TTSG-Y----G------
  PEPTIDE  ..................................................SG-Y----G------

                70                  90                  110
Amb a I.3  DWENNRQALADCAQGFAKGTYGGKWGDVYTVTSNLDDDVANPKEGTLRFAAAQNRPLWII
Amb a I.2  --A--------------------H--------DK-------------------------
Amb a I.1  --AE--K---------G---V---D--I-----E--------------G---------
  PEPTIDE  --AE--K---------G---V---D--I-u-..........................

                130                 150                 170
Amb a I.3  FKNDMVINLNQELVVNSDKTIDGRGVKVEIINGGLTLMNVKNIIIHNINIHDVKVLPGGM
Amb a I.2  --RN---H--------------------N-V-A-------------------I--C----
Amb a I.1  -ER----R-DK-M-----------A-------A-F--NG---V------M-----N---L
  PEPTIDE  ..................................................M-----u---L

                190                 210                 230
Amb a I.3  IKSNDGPPILRQASDGDTINVAGSSQIWIDHCSLSKSFDGLVDVTLGSTHVTISNCKFTQ
Amb a I.2  ----------Q----A-----------------AS---L-I----S---V-------
Amb a I.1  -------AAP-AG----A-SIS---------------V-----AK--T-RL-V--SL---
  PEPTIDE  --u-u-..........................................................

                250                 270                 290
Amb a I.3  QSKAILLGADDTHVQDKGMLATVAFNMFTDNVDQRMPRCRFGFFQVVNNNYDRWGTYAIG
Amb a I.2  HQFVL---------Y----------------H----------------------------
Amb a I.1  HQFVL-F--G-ENIE-R---------T-------------H-----------K--S----
  PEPTIDE  ..................................-----H----------K--S----

                310                 330                 350
Amb a I.3  GSSAPTILCQGNRFLAPDDQIKKNVLARTGTGAAESMAWNWRSDKDLLENGAIFVTSGSD
Amb a I.2  --------S-----F----I----------N----S----T-R---------LP----
Amb a I.1  --AS----S-----C---ERS-----G-H-EA-----K----TN--V--------A--V-
  PEPTIDE  --AS---.....................................................-

                370                 390
Amb a I.3  PVLTPVQSAGMIPAEPGEAAIKLTSSAGVFSCHPGAPC
Amb a I.2  -----E-K----------VLR-------L---Q----
Amb a I.1  -----E----------S-LS-------L--Q-----
  PEPTIDE  -----E----------S-LS-------Lu-Q-----
```

FIG. 2. **Comparison of the deduced amino acid sequences of clones *Amb a I.3, Amb a I.2,* and *Amb a I.1* peptides.** *Horizontal bars* (– – –) represent identity to clone *Amb a I.3*; a *dotted line* (·····) indicates no sequence information. *Asterisks* (*) indicate deletions in the amino acid sequences. Unidentified amino acid residues from the peptide sequencing are reprsented by a *u*.

performed using programs described previously (37, 38).

The presence of multiple naturally occurring forms of *Amb a I* in pollen was confirmed by two-dimensional electrophoresis of pollen protein extract (Fig. 3). Pollen proteins were separated first by charge and then by size, blotted onto nitrocellulose, and probed with a goat anti-*Amb a I* antiserum. Three major and three minor variants of *Amb a I* at 38 kDa could be detected with isoelectric points of approximately 5.2, 4.9, 4.7, 4.3, 5.9, and 5.4 in order of decreasing intensity. The calculated (from amino acid composition) isoelectric point of 5.2 of the clone representing *Amb a I.1,* along with its perfect match to *Amb a I* peptides, might suggest that this clone represents the major variant of the allergen isolated from pollen collected over a wide geographic area. The DNA and protein sequences were compared with sequences in the EMBL and NBRF computer data bases. No significant homologies were found in either case.

Northern blot analysis of RNA from defatted ragweed and meadow fescue (grass) pollens is depicted in Fig. 4. RNA was probed with a radiolabeled member of the *Amb a I.2* group. Strong binding can be seen to a 1.5-kilobase message in the ragweed pollen RNA, and no binding is demonstrable in comparable amounts of RNA isolated from the grass pollen. Similar analyses using RNA derived from other ragweed plant tissues such as leaves and roots also showed no binding to the *Amb a I* probe (data not shown).

A Southern blot of genomic DNA digested with various restriction enzymes and probed with *Amb a I.2* is shown in Fig. 5. The clone binds to multiple bands in all the preparations and, when compared with restriction enzyme maps of the clones, suggests that there are multiple genes encoding *Amb a I* in an outbred population of ragweed plants. Whether

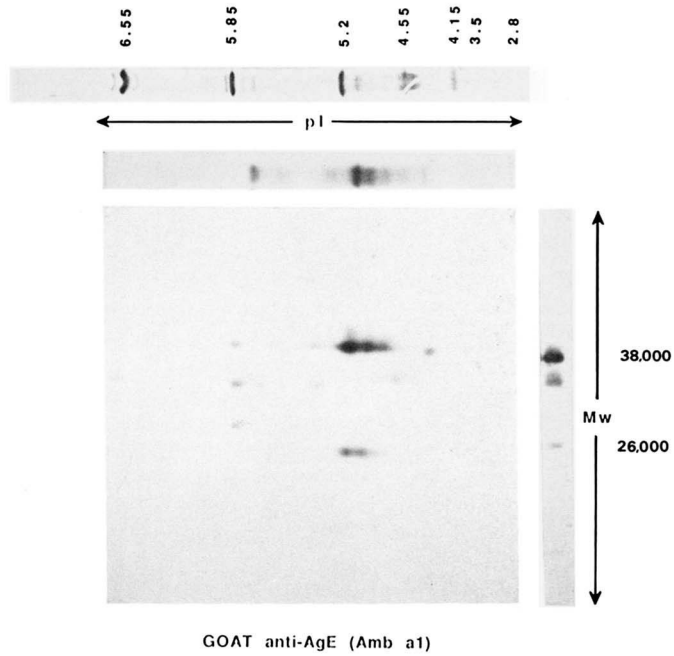


FIG. 3. **Two-dimensional gel electrophoresis and Western blotting of pollen protein.** Crude, soluble pollen proteins were subjected to isoelectric focusing (*left* to *right*) followed by 10% SDS-PAGE electrophoresis (*top* to *bottom*). The proteins were blotted onto nitrocellulose and probed with goat anti-*Amb a I* antisera. The 26,000-dalton band presumably represents the α chain of *Amb a I*.

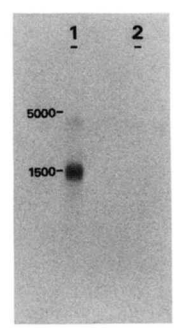


FIG. 4. **RNA blot analysis of total RNA from various tissues.** 20 μg of total RNA was electrophoresed, blotted onto nitrocellulose, and hybridized to radiolabeled *Amb a I.2. Lane 1,* short ragweed pollen; *lane 2,* meadow fescue pollen.

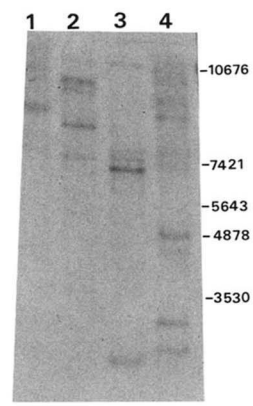


FIG. 5. **DNA blot analysis of short ragweed genomic DNA.** Genomic DNA was digested with *Bam*HI (*lane 1*), *Eco*RI (*lane 2*), *Hin*dIII (*lane 3*), and *Nco*I (*lane 4*) and run on a 0.7% agarose gel. The DNA was transferred to nitrocellulose and hybridized to radiolabeled *Amb a I.2.*

the observed heterogeneity of *Amb a I* is due to allelic variation, multiple genes in a single plant, or perhaps both is presently under investigation.

DISCUSSION

The use of molecular biology techniques to study directly the structure-function relationships of allergens of medical importance has only recently begun to be exploited. Thus far three non-plant allergens (white-faced hornet venom antigen 5, bee venom phospholipase A2, and the house dust mite allergen *Der p I*) have been cloned and used for immunological studies (39–41). In this report, the major allergen of short ragweed pollen has been cloned, and its complete nucleic acid and deduced amino acid sequence are described. This allergen (*Amb a I*) is shown to be a family of proteins, closely related to each other but unique from other families of proteins.

Proteins in the pollen coat can originate from two different sources. They may be expressed in the pollen cytoplasm itself, or they can be synthesized in the tapetal nursing cells surrounding the pollen and then deposited on the pollen surface (42). cDNA libraries, therefore, were made from poly(A^+) RNA both from maturing flowers and from commercially available defatted ragweed pollen. The pollen cDNA library was made in λgt11 and was screened with a panel of monoclonal antibodies raised specifically against denatured *Amb a I*. From the information gained by cloning and sequencing *Amb a I* from this library, it was possible to probe a flower cDNA library made in λgt10. In addition, once sequence information was derived, it became possible to design oligonucleotide primers to amplify cDNA coding specifically for *Amb a I* from ragweed flowers by means of the PCR reaction.

Initially, a pool of seven monoclonal antibodies raised against denatured *Amb a I* (recognizing both α and β chains) was used to screen the pollen cDNA library. Three clones were isolated and fully characterized. One of these, *Amb a I.2*, contained an almost complete coding sequence for *Amb a I*, another named *Amb a I.3* contained approximately two-thirds of the coding sequence, and the third, *Amb a I.1*, contained 320 base pairs and coded for only 50 COOH-terminal amino acids of *Amb a I*. Further screening of the libraries (and PCR amplification of cDNA in the case of *Amb a I.1*) has produced full-length sequences of three groups of highly related structures representing *Amb a I*. A total of 14 clones were completely sequenced. On average, clones within each group differed at less than 1% of the nucleotides coding for the structural protein while the difference between groups was of the order of 10–15% at the nucleotide level. When full-length representative clones from each group are expressed in *Escherichia coli* and the products blotted onto nitrocellulose membranes for probing with monoclonal antibodies to denatured *Amb a I*, each gives a unique pattern of reactivity with individual antibodies (data not shown). This suggests that individual antibodies are able to discriminate between clones and demonstrates that all three groups of *Amb a I* proteins were in the chromatographically prepared *Amb a I* used to originally immunize the mice. These three groups of *Amb a I* clones, therefore, are expressed and are not transcripts of nonfunctional genes.

Multiple isoelectric forms and conservative amino acid substitutions are a common feature of pollen allergens that have been well characterized such as those from the grasses and from short ragweed (4, 43). It was shown earlier (4, 6) that *Amb a I* has four electrophoretic and/or structural forms called A, B, C, and D. These forms are indistinguishable by amino acid compositional analysis and comparison of their antigenic and allergenic properties. In this report, two-dimensional Western analysis of pollen extracts shows three major isoelectric forms of *Amb a I* and a number of minor forms that differ both in size and charge. This is consistent with the description of three groups of *Amb a I* clones and the appreciation that individual members of a group may differ by as much as 1% from the prototype within a group. Current efforts are directed at assigning specific groups to specific isoelectric forms of *Amb a I*. It is not clear at this time that the A, B, C, and D forms of *Amb a I* have any direct relationship to *Amb a I.1*, *Amb a I.2*, or *Amb a I.3* as described in this report.

Since the pollen used to create the cDNA library was collected from a wide geographic area, it is unlikely that other major *Amb a I* sequences will be found in short ragweed. Current experiments are under way to determine whether individual short ragweed plants collected from various geographical locations in North America express multiple (or all) members of the *Amb a I* family or whether individual plants show some restricted *Amb a I* production. Southern blot analysis of genomic DNA clearly shows multiple restriction fragments identified with a radiolabeled *Amb a I* probe, and it will be of interest to determine, in individual plants, whether allelic variation, multiple structural genes, or both contribute to this phenomenon.

Amino acid sequence analyses of peptides from chromatographically isolated *Amb a I* have been obtained. It is of great interest that the best match of that amino acid sequence is with the *Amb a I.1* group of *Amb a I* sequences. However, despite exhaustive screening of several independently derived cDNA libraries, only one short (320-base pair) clone belonging to this group was ever isolated, making this clone the most underrepresented member of the *Amb a I* family. A full-length clone of *Amb a I.1* was only obtained by sequence-specific amplification of pollen cDNA using primers specific for the 3′ end of the original 320-base pair clone and sequences shared between *Amb a I.2* and *Amb a I.3* at the 5′ end. Three possible explanations for this enigma are currently being considered. First, if it is the most actively transcribed and/or translated message, it could have a high turnover rate *in vivo* or be particularly prone to degradation during storage or processing of the pollen. Second, it is possible that its structure could somehow interfere uniquely with cDNA synthesis and cloning efficiency, and its paucity in the libraries merely represents a technical difficulty. Third, the *Amb a I.1* protein could in fact be a minor *Amb a I* variant but be selectively enriched for during the purification of *Amb a I* from pollen extract.

The amino acid composition of the cloned proteins corresponds well to experimental results reported previously (4). The presence of cysteine as the carboxyl-terminal deduced amino acid in all cDNA clones studied is in contrast to a report in which carboxypeptidase digestion of chromatographically purified *Amb a I* showed that leucine was the carboxyl-terminal amino acid (6). To settle the question of the carboxyl-terminal residue, the aspartic acid-proline peptide bond (position 360–361, Fig. 2) of chromatographically purified *Amb a I* was cleaved by 70% formic acid. The peptide sequence starting with proline was obtained in the digestion mixture whereas the nonspecific partially cleaved protein background sequence was suppressed by *o*-phthalaldehyde treatment. The 38-residue peptide sequence obtained corresponded completely with the deduced sequence of *Amb a I.1* up to and including the terminal cysteine.

Since the amino terminus of *Amb a I* is "blocked" (6) and not amenable to direct Edman degradation, it is important to be able to conclude that the clones reported here contain the full coding sequence for *Amb a I*. Two lines of evidence support that conclusion. The experimental evidence comes from a

previous study (44) in which primer extension was used to add nucleotides to the 3′ end of an oligonucleotide designed from a peptide sequence using flower mRNA as template. The 5′ end of the oligonucleotide corresponds to nucleotide 249 in clone *Amb a I.3* and produced a cDNA fragment about 250 nucleotides long, suggesting that very little 5′-untranslated sequence is missing from the full-length message. The second piece of evidence that these are full-length clones coding for *Amb a I* comes from a study of the area around the presumed translation initiation codon AUG (45, 46). From a study of 211 eucaryotic messages, a consensus sequence has been identified at the initiation site. The sequence contains a purine (preferably an A) at position −3, a G at position +4, and a predominance of C at positions −1, −2, −4 and −5. Clone *Amb a I.3* has A at −3, G at +4, and C at −2.

There is an obvious hydrophobic stretch of amino acids following the initiating methionine, and this area has been examined to predict the cleavage site between the hydrophobic leader and the secreted protein. The prediction is based upon data collected for 450 secreted eucaryotic proteins (47) with known cleavage sites. Positions −1 and −3 are most critical for signal peptide cleavage, and the alanine at position 26 in *Amb a I* best fulfills the criteria associated with the cleavage site. For instance, if alanine 26 is indeed residue +1, then the −1 and −3 positions are serine and valine, respectively. The frequency of those amino acids in those positions in the data base is two to three times the expected values if random amino acids occupied those positions. All amino acids, from −1 through −13, are consistent with the tabulations from the 450 known cleavage sites. Current efforts are under way to address directly the question of how and where the *Amb a I* proteins are blocked.

At the 3′ end, the untranslated region does not have the AATAAA polyadenylation signal observed in most animal sequences located 9–23 bases upstream from the poly(A) tail. However, there is a variant sequence, AATGAA, located 45 bases from the poly(A) addition site, which might serve as the adenylation signal; and another variation of this sequence, AAAAAT, is located 39 bases downstream from the TAA stop codon. A similar sequence, AATAAT is located 101 nucleotides upstream from the polyadenylation site in clone *Amb a I.2*. Poly(A) signal patterns are known to be much more complex in plants than animals (48) in their sequence, number, and locations. For example, a pollen-specific sequence from maize, Zmo13, reported by Hanson *et al.* (49), contains the consensus motif AATAAA 180 bases upstream from the poly(A) site and two variants, AATATA and AATTAT, located at 55 and 44 bases respectively. Another pollen-expressed gene, alcohol dehydrogenase, has the AATTAT sequence centered 44 bases upstream from its polyadenylation site.

The data presented in this paper firmly establish that *Amb a I* (antigen E) is a family of closely related proteins. Experiments are currently in progress to address the relative abundance of each family member in pollen, as mentioned above. Just as importantly, it is necessary to determine the immunogenicity and allergenicity of the *Amb a I* family members. It has been possible to express prototypic clones for each family member in procaryotic cells.[3] This material is being used to compare T cell proliferation and IgE binding from individual ragweed allergic patients. This information can potentially be used to devise an effective course of ragweed immunotherapy.

[3] J. Bond, R. Garman, K. Keating, T. Briner, T. Rafnar, D. Klapper, and B. Rogers, manuscript in preparation.

Acknowledgments—We wish to thank Drs. Joseph Olson, James Smith, and Robert Esch for providing reagents; Dr. David Marsh for providing goat anti-*Amb a I* IgG; Tom Wells, Anneliese Nault, Robert Schreifels, Andrew Brauer, and Joanne Pollock for technical assistance; and Joyce Bradshaw for preparing the manuscript.

REFERENCES

1. Marsh, D. G., Goodfriend, L., King, T. P., Lowenstein, H., and Platts-Mills, T. A. E. (1986) *Bull. WHO* **64,** 767–770
2. King, T. P. (1976) *Adv. Immunol.* **23,** 77–105
3. Løwenstein, H., and Marsh, D. G. (1983) *J. Immunol.* **130,** 727–731
4. King, T. P., Norman, P. S., and Connell, J. T. (1964) *Biochemistry* **3,** 458–468
5. Zeiss, C. R., Pruzansky, J. J., Patterson, R., and Roberts, M. (1973) *J. Immunol.* **110,** 414–421
6. King, T. P., Alagon, A., Kochoumian, L., Kuan, J., Sobotka, A. K., and Lichtenstein, L. M. (1981) *Arch. Biochem. Biophys.* **212,** 127–135
7. Norman, P. S., Winkenwerder, W. L., Lichtenstein, L. M., and Tignall, J. (1968) *J. Allergy* **42,** 93–108
8. Norman, P. S., King, T. P., Alexander, J. F., Kagey-Sobotka, A., and Lichtenstein, L. M. (1984) *J. Allergy Clin. Immunol.* **73,** 782–789
9. Butterfield, J. H., Gleich, G. J., Yunginger, J. W., Zimmerman, E. M., and Reed, C. E. (1981) *J. Allergy Clin. Immunol.* **67,** 272–278
10. Zeiss, C. R., Metzgar, W. J., and Levitz, D. (1977) *Clin. Exp. Immunol.* **28,** 250–255
11. Grammer, L. C., Zeiss, C. R., Suszko, I. M., Shaughnessy, M. A., and Patterson, R. (1982) *J. Allergy Clin. Immunol.* **69,** 494–499
12. Olson, J. R., and Klapper, D. G. (1986) *J. Immunol.* **136,** 2109–2115
13. DeLisi, C., and Berzofsky, J. A. (1985) *Proc. Natl. Acad. Sci. U. S. A.* **82,** 7048–7052
14. Ishizaka, K., Kishimoto, T., Delespesse, G., and King, T. P. (1974) *J. Immunol.* **113,** 70–77
15. Ishizaka, K., Okudaira, H., and King, T. P. (1975) *J. Immunol.* **114,** 110–115
16. Djurup, R. (1985) *Allergy* **40,** 469–486
17. Smith, J. J., Olson, J. R., and Klapper, D. G. (1988) *Mol. Immunol.* **25,** 355–365
18. Rivin, C. J., Zimmer, E. A., and Walbot, V. (1982) in *Maine for Biological Research* (Sheridan, W. F., ed) Plant Molecular Biology Association, Charlottesville, VA
19. Lagrimini, L. M., Burkhart, W., Moyer, M., and Rothstein, S. (1987) *Proc. Natl. Acad. Sci. U. S. A.* **84,** 7542–7546
20. Levy, S., Mendel, E., and Kon, S. (1987) *Gene (Amst.)* **54,** 167–174
21. Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual,* pp. 97–148, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
22. Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–5467
23. Gubler, U., and Hoffman, B. J. (1983) *Gene (Amst.)* **25,** 263–269
24. Huyuk, T. V., Young, R. A., and Davis, R. W. (1985) in *DNA Cloning: A Practical Approach* (Glover, D., ed) Vol. 1, pp. 49–78, IRL Press, Oxford
25. Benton, W. D., and Davis, R. W. (1977) *Science* **196,** 180–182
26. Rigby, P. W. J., Dieckmann, M., Rhodes, C., and Berg, P. (1977) *J. Mol. Biol.* **113,** 237–251
27. Matsudaira, P. (1987) *J. Biol. Chem.* **262,** 10335–10038
28. Andrews, P. C., and Dixon, J. E. (1987) *Anal. Biochem.* **161,** 524–528
29. Terhorst, C., Robb, R., Jones, C., and Strominger, J. L. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74,** 4002–4006
30. Brauer, A. W., Oman, C. L., and Margolies, M. N. (1984) *Anal. Biochem.* **137,** 134–142
31. Jungföretagenab, O. (ed) (1982) *Isoelectric Focusing: Principles and Methods,* pp. 127–168, Pharmacia/LKB Biotechnology Inc., Piscataway, NJ
32. Celis, J. E., and Bravo, R. (eds) (1984) in *Two-dimensional Gel Electrophoresis of Proteins: Methods and Applications,* Academic Press, New York
33. Towbin, H., Staehelin, T., and Gordon, J. (1979) *Proc. Natl. Acad. Sci. U. S. A.* **76,** 4350–4354

34. Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J., and Rutter, W. J. (1979) *Biochemistry* **18,** 5294–5299
35. Saiki, R. K., Scharf, S., Falodna, F., Mullis, K. B., Horn, G. T., Erlich, H. A., and Arnheim, N. (1985) *Science* **230,** 1350–1354
36. Roux, K. H., and Dhanarahan, P. (1990) *Biotechniques* **8,** 48–57
37. Keller, C., Corcoran, M., and Roberts, R. J. (1984) *Nucleic Acids Res.* **12,** 379–386
38. Devereux, J., Haeberli, P., and Smithies, O. (1984) *Nucleic Acids Res.* **12,** 387–395
39. Fang, K. S. Y., Vitale, M., Fehlner, P., and King, T. P. (1988) *Proc. Natl. Acad. Sci. U. S. A.* **85,** 895–899
40. Thomas, W. R., Stewart, G. A., Simpson, R. J., Chua, K. Y., Plozza, T. M., Dilworth, R. J., Nisbet, A., and Turner, J. K. (1988) *Int. Arch. Allergy Appl. Immunol.* **85,** 127–129
41. Kuchler, K., Gmachl, M., Sippl, M., and Kreil, G. (1989) *Eur. J. Biochem.* **184,** 249–254
42. Knox, R. B. (1979) in *Pollen and Allergy,* pp. 56–57, University Park Press, Baltimore
43. Marsh, D. G. (1975) in *The Antigens* (Sela, M. ed) Vol. 3, pp. 271–359, Academic Press, New York
44. Klapper, D. G., Woods, S., Olson, J., Esch, R., Smith, J. J., and Rafnar, T. (1989) in *Advances in the Biosciences* (Said El Shami, A., and Merrett, T. G., eds) vol. 74, pp. 149–159, Pergamon Press, Elmsford, NY
45. Kozak, M. (1981) *Nucleic Acids Res.* **9,** 5233–5252
46. Kozak, M. (1984) *Nucleic Acids Res.* **12,** 857–872
47. von Heijne, G. (1986) *Nucleic Acids Res.* **14,** 4683–4690
48. Lycett, G. W., Delauney, A. J., and Croy, R. R. D. (1983) *FEBS Lett.* **153,** 43–46
49. Hanson, D. D., Hamilton, D. A., Travis, J. L., Bashe, D. M., and Mascarenhas, J. P. (1989) *Plant Cell* **1,** 173–179